第 34 卷 第 8 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.8 Computer Engineering April 2008
·安全技术· 文章编号：1000—3428(2008)08—0184—03 文献标识码：A 中图分类号：TP391

基于核密度估计的入侵检测方法

周 璨[1]，李伯阳[2]，黄 斌[2]，刘 刘[2]
(1. 衡阳师范学院数学系，衡阳 421008；2. 厦门大学软件学院，厦门 361005)

摘 要：通过分析现有入侵检测技术的不足，探讨基于孤立点挖掘的入侵检测技术的优势，提出一种基于核密度估计的入侵检测方法。该方法通过核密度估计求出孤立点的近似集，再通过筛选近似集获得最终的孤立点集合，从而检测入侵记录。阐述了具体实现方案，通过仿真实验验证了该方法的可行性。

关键词：入侵检测；孤立点；核密度估计；编码映射；主成分分析

Intrusion Detection Method Based on Kernel Density Estimator

ZHOU Can[1], LI Bo-yang[2], HUANG Bin[2], LIU Liu[2]
(1. Department of Mathematics, Hengyang Normal University, Hengyang 421008; 2. Software School, Xiamen University, Xiamen 361005)

【Abstract】This paper analyses the disadvantages of the existing intrusion detection technology and discusses the advantages of intrusion detection based on outlier mining, a new intrusion detection method based on kernel density estimator called IDKD is proposed. In IDKD, the approximate set of outliers is calculated by kernel density estimator through one data set pass, and the indeed set of outliers is generated from the approximate set by another data set pass, the anomaly records are detected. This method is applied in KDD99 data set and gets satisfactory results.

【Key words】intrusion detection; outlier; kernel density estimator; code mapping; principal components analysis

现有的入侵检测技术大致可以归为两类：有监督的入侵检测技术与无监督的入侵检测技术。有监督的入侵检测最具代表的是基于 SVM 和神经网络的入侵检测技术[1]。通常这类技术对于已知的攻击类型能达到很高的检测率但无法检测未知类型的攻击。无监督的入侵检测技术以基于聚类的入侵检测[2]为主，其中以 K-MEANS 算法运用最为广泛。无监督的入侵检测具有不需要训练、可以识别未知类型攻击的优点，同时存在误检率过高和不易处理符号型数据的缺点。与聚类分析不同，孤立点挖掘[3-4]作为数据挖掘领域中的一个重要研究方向，其任务在于从大量复杂的数据中挖掘出新颖的、与常规数据模式明显不同的异常数据模式。在入侵检研究领域中正常记录和入侵记录存在明显差别；且入侵记录的数量远低于正常记录的数量。这使得入侵记录完全符合孤立点的定义，因此，将入侵检测转化为孤立点挖掘比进行聚类分析更为合理，而且更能反映入侵行为的本质。本文基于核密度估计提出一种新的入侵检测方法 IDKD(Intrusion Detection Based on Kernel Density Estimator)，该方法通过扫描一遍数据集对数据点进行核密度估计得到孤立点候选集，再通过重新估计候选集数据的核密度得到修正后的孤立点近似集，最终对近似集筛选得到真正的孤立点集合，从而得出检测结果。

1 基于核密度估计的孤立点挖掘

本文采用文献[4]中基于距离的孤立点定义：

定义 令 x 为数据集 D 中的一个数据点，称 x 为 $D(p, r)$ 孤立点，当且仅当 D 中至多有 p 个点到 x 的距离不超过 r。

记 $N(x, r)$为以 x 为中心 r 为半径的区域内的数据点个数，则 x 为孤立点的充要条件为 $N(x, r)\leq p$。那么孤立点集合 $Out=\{x^i \mid N(x^i,r)\leq p, x^i\in D\}$。

由该定义描述的孤立点有如下 2 个性质：(1)孤立点的分布较正常点稀疏得多；(2)孤立点数量远小于正常点数量。这 2 个性质与关于入侵记录的 2 个事实一致，因此通过该定义来描述入侵记录在理论上是可行的。

1.1 多维核密度估计

核密度估计是一种基于核理论[5]的非参数估计方法。对于某一数据集 D，其多维核密度估计函数形式如下：

$$\hat{f}(x_1,x_2,\cdots,x_d)=\frac{1}{bh_1\cdots h_d}\sum_{x^i\in S}k(\frac{x_1-X_1^i}{h_1},\frac{x_2-X_2^i}{h_2},\cdots,\frac{x_d-X_d^i}{h_d}) \quad (1)$$

其中，k 为核函数；集合 S 为对数据集 D 进行随机抽样得到的样本集；参数 h_i 为不同维度上的 bandWidth，用于控制核估计函数的平滑度。

常用的核函数有很多，如高斯核函数和 Epanechnikov 核函数等。文献[5]研究表明，在观测样本和 bandWidth 相同的情况下，不同形状的核函数对密度估计的影响不大。考虑到易于积分和高效性，本文选择了 Epanechnikov 核函数，中心位于 X^i 的 Epanechnikov 核函数形式如下：

$$k(\frac{x_1-X_1^i}{h_1},\frac{x_2-X_2^i}{h_2},\cdots,\frac{x_d-X_d^i}{h_d})=$$
$$(3/4)^d\times\prod_{1\leq j\leq d}((1-(\frac{x_j-X_j^i}{h_j})^2(|\frac{x_j-X_j^i}{h_j}|\leq 1)) \quad (2)$$

其中，当条件 A 满足时，函数(A)=1，否则(A)=0。通过式(1)和式(2)可以估计出 D 中的任意数据点 x 处的概率密度。在以 x 为中心 $2r$ 为边长的超矩形区域内，通过对式(1)进行积分可以得到 $N(x, r)$的近似：

$$\hat{N}(x,r)=n(\tfrac{3}{4})^d\frac{1}{bh_1\cdots h_d}\times\sum_{1\leq i\leq b}\prod_{1\leq j\leq d}\int_{[o_j-r,o_j+r]}(1-(\frac{x_j-X_j^i}{h_j})^2)\mathrm{d}x_j \quad (3)$$

其中，x 的坐标由 $(o_1,o_2,\cdots,o_d)$ 表示。

作者简介：周 璨(1981－)，女，助教，主研方向：概率统计，模糊数学等；李伯阳、黄 斌、刘 刘，硕士研究生

收稿日期：2007-04-30 **E-mail：**i_liby@163.com

通过式(3)计算 D 中每一个数据点的 $\hat{N}(x,r)$ 值，可以得到孤立点候选集合 $\hat{Out}=\left\{x^i \mid \hat{N}(x^i,r)\le p \wedge x^i \in D\right\}$。

1.2 观测样本的选择

一个核密度估计函数的好坏取决于 2 个因素：(1)所选择的观测样本集合；(2)bandWidth 的取值。

通常情况下采取随机抽样的方式来选择观测样本，并且抽样的比例越大则得到估计函数越能够接近于真实的密度函数。但从算法时间效率的角度来讲，选择的观测样本越多，计算核密度估计函数花费的时间越长。

以孤立点挖掘为目的进行核密度估计旨在以最低的时间消耗发现数据集中的孤立点集，因此希望被选取的观测样本具有以下 2 种性质：(1)尽可能覆盖到数据集的整个范围；(2)尽可能选择局部密度大的数据点。第 1 个性质确保得到的核密度估计函数能全面地近似描述真实的概率密度分布；第 2 个性质则使得孤立点在计算核密度估计函数后得到的 $\hat{N}(x,r)$ 很小，而其他非孤立点计算得到的 $\hat{N}(x,r)$ 很大。因此，通过式(3)得到的孤立点候选集与真正的孤立点集合更为近似。

对于观测样本集 S 的选择，本文采用文献[6]提出的 Greedy 算法并进行适当的修改，在计算数据点到样本集的距离时加入数据点局部密度权重，其具体步骤如下：

(1)初始时 $S=\phi$;随机选择一个数据点作为第 1 个观测样本 X^1， $S=S\cup X^1$；

(2) 计算 $\{dist(x^i,S)\mid x^i\in D\wedge x^i\notin S\}$，$dist(x,S)=\min\limits_{x^i\in S}\{w(x)\|x,X^i\|\}$，其中，$w(x)$表示数据点 x 的局部密度权值， 当 $\min\limits_{x^i\in D\wedge x^i\notin S}\{dist(x^i,S)\}<\varepsilon$ 则停止并返回 S；否则就变成

$$S=S\cup\left\{x^j \mid dist(x^j,S)=\min_{x^i\in D\wedge x^i\notin S}\{dist(x^j,S)\}\right\}$$

(3)转移至(2)。

1.3 孤立点集合的确定

文献[3]中提出一种基于核密度估计的孤立点挖掘方法——在随机选择观测样本的基础上通过一次扫描数据集获得孤立点候选集 $\hat{Out}$，再通过对 $\hat{Out}$ 筛选获得真正的孤立点集 Out ——这种方法往往要求较大的观测样本集来保证密度估计的准确性，同时因为得到的 $\hat{Out}$ 通常较大而使得筛选过程时间消耗较高。本文通过改进的 Greedy 算法选取观测样本，旨在用尽可能少的观测样本获得尽可能准确的密度估计，从而进一步提高算法的时间效率。因此，本文通过如下方法确定孤立点集合：

在初始情况下，由于无法得知数据集的密度分布，设定权值 $w(x^i)=1,\forall x^i\in D$ 运行 Greedy 算法选择观测样本集 S。通过扫描数据集 D 计算 $\hat{N}(x^i,r),\forall x^i\in D$，更新权值 $w(x^i)=\hat{N}(x^i,r),\forall x^i\in D$ 并获得孤立点候选集 $\hat{Out}$。基于新的权值运行 Greedy 选择新的 S，重新计算 $\hat{N}(x^i,r),\forall x^i\in\hat{Out}$ 对候选集 $\hat{Out}$ 进行修正，得到与 Out 更近似的集合 $\tilde{Out}=\{x^i\mid\hat{N}(x^i,r)\le p\wedge x^i\in\hat{Out}\}$。最后，计算 $N(x^i,r),\forall x^i\in\tilde{Out}$，得到孤立点集合 Out。

1.4 bandWidth 的选择

本文采用 LCSV 方法[5]确定 bandWidth $H=(h_1,h_2,\cdots,h_d)$ 的取值。LCSV 是基于积分平方误差(Integrated Square Error)最小准则的一种计算方法，其公式如下：

$$LCSV(H)=\frac{1}{n^2h_1\cdots h_d}\sum_{i=1}^{b}\sum_{j=1}^{b}\bar{k}(X^i-X^j)-\frac{2}{n(n-1)h_1\cdots h_d}\sum_{i=1}^{b}\sum_{j\ne i}^{b}k\left(\frac{X^i-X^j}{H}\right)\quad(4)$$

其中，$\bar{k}(v)=\int k(u)k(v-u)\mathrm{d}u$ ；当 $LCSV(H)$取最小值时，H 即为所求。

2 基于核密度估计的入侵检测方法 IDKD

如图 1 所示，基于核密度估计的入侵检测流程分为 4 步：

(1)对网络连接数据进行特征选择，通过极差标准化将数据空间转换到 $[0,1]^d$ ；

(2)通过 Greedy 算法选择观测样本同时计算 bandWidth，扫描数据集 D，计算 $\hat{N}(x^i,r),\forall x^i\in D$，得到每个 x^i 对应的 w^i 和孤立点候选集 $\hat{Out}$ ；

(3)通过局部密度加权的 Greedy 重新选择观测样本，扫描候选集 $\hat{Out}$，计算 $\hat{N}(x^i,r),\forall x^i\in\hat{Out}$，得到孤立点近似集 $\tilde{Out}$ ；

(4)扫描 $\hat{Out}$，计算 $N(x^i,r),\forall x^i\in\tilde{Out}$，获得孤立点集合 Out，输出检测结果。

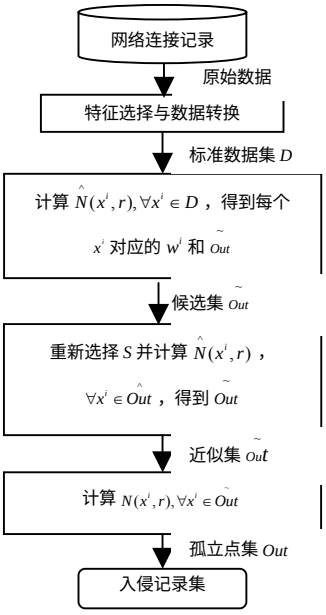

图 1 基于核密度估计的入侵检测流程

本文采用的实验数据集来自于 KDD Cup 1999，该数据集提供了 4 900 000 条连接数据，每个连接共有 41 个属性，其中有 8 个属性是离散型，其余是连续型，最后一个属性标示该连接记录是否为入侵记录。KDD99 数据集中主要包含了 4 大类入侵类型，即拒绝服务攻击、远程计算机的未授权访问、权限提升和漏洞探测。其中，DoS 攻击所占比例最大。

在这 41 维属性中，不同的属性对于入侵检测有着不同的重要性。文献[1]通过支持向量机的方法对 KDD Cup 1999 数据集进行降维，得出最重要的 13 个属性，包括 11 个数字型属性 duration, src_bytes, dst_bytes, urgent, count, srv_count, same_srv_rate, dst_host_count, dst_host_srv_count, dst_host_same_src_port_rate 和 dst_host_same_srv_rate，与 2 个符号型属性 protocol_type 和 service。

本文选取上述 13 个属性进行实验分析，并对其中 2 个符号属性采用编码映射方法进行展开，同时对编码映射后的数据集进行极差标准化。鉴于编码映射后数据维度的剧增，本文对扩展后的数据集进行主成分分析(PCA)，通过选取得分最高的 3 到 5 个主成分重构该数据集，最终达到降维的效果。

3 实验结果分析

实验从 3 个方面检验基于核密度估计的入侵检测方法 IDKD 的性能：(1)真实网络环境下的性能；(2)对于不同类型攻击的检测效果；(3)IDKD 对于大小不同的数据集的稳定性。为此，实验抽取了 4 个混合攻击数据集和 4 个单一不同类型攻击记录集，其中，正常记录与攻击记录比例为 50:1。8 个数据集大小不一，具体如表 1 所示。

表 1　测试数据集明细表

数据集	正常记录	入侵记录	总数	类型
Mix1	9 800	200	10 000	Mixed
Mix2	19 600	400	20 000	Mixed
Mix3	29 400	600	30 000	Mixed
Mix4	39 200	800	40 000	Mixed
PROBE	43 853	835	44 688	PROBE
R2L	19 600	365	19 965	R2L
U2R	29 400	586	29 986	U2R
DOS	48 613	781	49 394	DOS

K-MEANS 算法和 IDKD 方法对于 8 个测试数据集的检测率和误检率由表 2 列出。

表 2　测试结果明细表　(%)

数据集	K-MEANS		IDKD	
	检测率	误检率	检测率	误检率
Mix1	91.6	1.93	94.2	1.71
Mix2	88.3	2.36	91.3	2.03
Mix3	80.9	4.11	88.9	2.35
Mix4	86.7	2.13	90.0	1.90
PROBE	95.4	5.76	97.0	2.09
R2L	41.3	21.20	50.1	13.20
U2R	27.7	24.60	28.4	11.50
DOS	93.5	1.69	98.1	0.87

对于 PROBE 和 DoS 类型的攻击，IDKD 方法取得了较高的检测率和较低的误检率。这是由于 PROBE 和 DoS 两类攻击的模式比较单一，在入侵记录中所占的比例较大而且与正常行为有明显的差异。而对于 R2L 和 U2R，IDKD 方法无论在检测率还是误检率都表现很差。其原因在于 R2L 和 U2R 往往只存在于非常少的连接当中，难以对其进行足够的特征刻画，因此，在 PCA 降维过程中与其相关的主成分容易因为得分低而被忽略，从而丢失对这类数据的孤立性的描述。

对于模拟真实网络环境的混合攻击测试，IDKD 方法的检测率和误检率分别稳定在 90%和 2%左右。这与 IDKD 在单一攻击的的表现基本吻合，当混合攻击中的 R2L 和 U2R 较少时，检测效果较好，而一旦 R2L 和 U2R 增多，检测效果会随之下降。从表 2 中与 K-MEANS 的结果对比看出，无论是对于混合型的攻击还是针对某一类特定的攻击，IDKD 方法的检测率和误检率都优于 K-MEANS 算法，检测性能也更加稳定。

4 结束语

本文分析了孤立点与网络入侵行为在统计学上的相似性，提出一种基于核密度估计的入侵检测方法 IDKD，该方法通过 3 次扫描确定记录集中的孤立点，从而完成对入侵行为的检测。通过在 KDD Cup99 上进行仿真实验，验证了该方法对于接近于真实网络环境下的混合攻击有较为稳定和有效的检测结果，尤其对于 DoS 和 PROBE 两类攻击 IDKD 达到了较高的检测率和较低的误检率。下一步的研究工作将着力于提高对 U2L 和 R2L 类型攻击的检测效果。

参考文献

[1] Mukkamala S, Janoski G, Sung A H. Intrusion Detection Using Support Vector Machines and Neural Networks[C]//Proc. of the IEEE Int'l Joint Conf. on Neural Networks. [S. l.]: IEEE Press, 2002: 1702-1707.

[2] PortnoyL, Eskin E, Stolfo S J. Intrusion Detection with Unlabeled Data Using Clustering[C]//Proceedings of ACM CSS Work-shop on Data Mining Applied to Security. Philadelphia, PA, USA: ACM Press, 2001.

[3] Kollios G, Gunopoulos D, Koudas N, et al. Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Datasets[J]. IEEE Trans. on Knowledge Data Eng., 2003, 15(5): 1170-1187.

[4] Knorr E, Ng R. Algorithms for Mining Distance Based Outliers in Large Databases[C]//Proc. of Very Large Databases Conf.. New York, USA: Morgan Kanfmann, 1998: 392-403.

[5] Scott D. Multivariate Density Estimation: Theory, Practice and Visualization[M]. [S. l.]: Wiley and Sons, 1992.

[6] Gonzalez T. Clustering to Minimize the Maximum Intercluster Distance[J]. Theoretical Computer Science, 1985, 38(2): 293-306.

(上接第 183 页)

图 2 表示了路长 L 取不同值时，能正确找到发送者的概率 P_s 与观测次数 n 的关系。

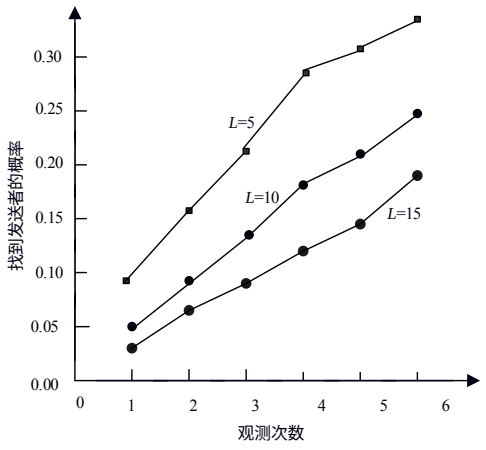

图 2　发送者概率与观测次数的关系

由图 2 可知，若路长固定，攻击者成功找到发送者的概率随着观测次数的增加而增大。随着路长的增大，成功找到发送者概率 P_s 增加缓慢。在相同观测次数下，L 值越小，攻击者成功找到发送者的概率越大，也就是说，成功找到发送者概率 P_s 的值随着路长 L 的增大而迅速下降。当路长选取适当时，不能找到发送者的概率将变大，这样就实现了高匿名度的效率。

4 结束语

本文提出了一个基于重路由技术的匿名通信系统模型，基于该模型，给出了一种攻击模型及攻击算法，实验表明了该算法的有效性。

参考文献

[1] 徐红云, 陈建二. 匿名通信系统中统计型攻击模型研究[J]. 小型微型计算机系统, 2004, 25(11): 1926-1929.

[2] Newman-Wolf R E. Venkatraman B R. Performance Analysis of a Method for High Level Prevention of Traffic Analysis[C]//Proc. of the 10th Annual Computer Security and Application Conference. Orlando, Florida, USA: [s. n.], 1994-10: 5-9.